MS-A0503 First course in probability and statistics

2A Expected value and transformations

Jukka Kohonen

Deparment of mathematics and systems analysis
Aalto SCI

Academic year 2020–2021
Period III

Contents

Expected value

The expected value (or expectation or mean) of a discrete random number X is

$$\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x) = \sum_x x\, f(x)$$

where the sum is taken over the possible values of X.

Or: It is the *probability-weighted average of the possible values*.

Example (Die result)

The expected value of one die rolled is

$$\mathbb{E}(X) = (1 \times \frac{1}{6}) + (2 \times \frac{1}{6}) \cdots + (6 \times \frac{1}{6}) = 3.5.$$

What does $\mathbb{E}(X)$ tell about the random variable?
(The name is misleading. It is not really a value that is "expected" to occur, because the die result is never 3.5.)

Expected value vs. long-term average

Let us play n rounds of a game, where each round gives a random payoff of X. With density $f(x) = \mathbb{P}(X = x)$.

Suppose that the payoff x occurs **approximately** $n\,f(x)$ times.

- Then our *total* payoff is approximately

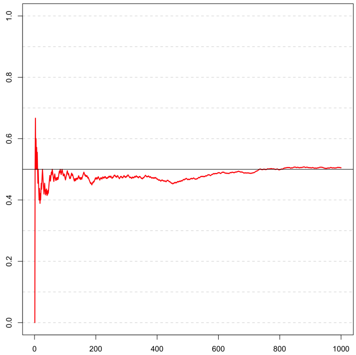
$$\sum_x x\,n\,f(x).$$

- Then our *average-per-round* payoff is approximately

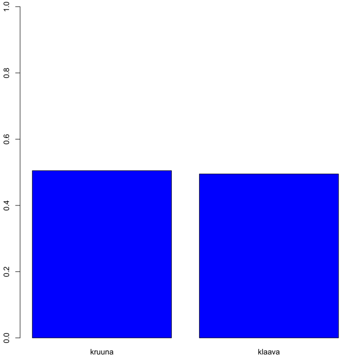
$$\frac{1}{n}\sum_x x\,n\,f(x) \;=\; \sum_x x\,f(x) \;=\; \mathbb{E}(X).$$

But is the thing true that we supposed?

Example: 1000 coin tosses



Relative frequency of heads, as number of tosses grows

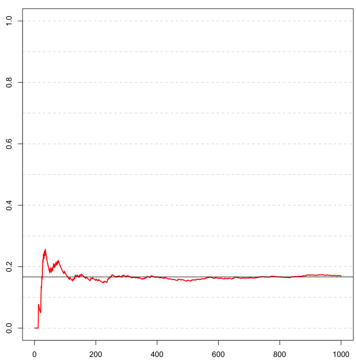


Relative frequencies of heads and tails after 1000 tosses.

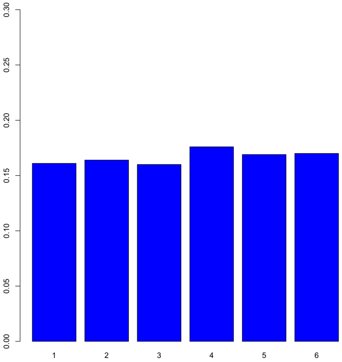
```
n <- 1000
x <- sample(c(0,1),n,replace=TRUE)
plot(cumsum(x)/(1:n),type="l")
plot(table(x))
```

http://www.r-project.org/
http://www.random.org/

Another example: 1000 rolls of a die



Relative frequency of sixes as number of rolls grows



Relative frequencies of each value, after 1000 rolls

```
n <- 1000
x <- sample(1:6,n,replace=TRUE)
plot(cumsum(x==6)/(1:n),type="l")
plot(table(x))
```

http://www.r-project.org/
http://www.random.org/

Contents

Variables: Realized average is near the expected value

Proposition (Law of large numbers (LLN))

If $X_1, X_2, X_3, \ldots$ are independent random numbers, and each has the same distribution as random number X, then the event

$$\frac{1}{n}\sum_{i=1}^{n} X_i \;=\; \mathbb{E}(X) \pm 0.001$$

is true with probability that approaches 1, as n grows.

This is the fundamental theorem of stochastics. Basically, it says the *randomness of the average gradually disappears* as n grows.

- The average $\frac{1}{n}\sum_{i=1}^{n} X_i$ is a random number
- The expectation $\mathbb{E}(X)$ is a deterministic, single number
- In place of 0.001, you can put any number $\epsilon > 0$.

Does it hold if the X_i are *dependent* (e.g. consecutive rainfalls)?
Not necessarily, but yes if the dependence is weak enough (ergodicity).

Events: Realized frequency is near the probability

Proposition

If $X_1, X_2, \ldots$ are independent random variables distributed like X, then for any set B of possible values, the relative frequency of B in the sequence $(X_1, \ldots, X_n)$ fulfills

$$\frac{\#\{i \in \{1, 2, \ldots, n\} \ : \ X_i \in B\}}{n} \ = \ \mathbb{P}(X \in B) \pm 0.001$$

with a probability approaching 1 as n grows.

- Example: Because density at x is $f(x) = \mathbb{P}(X = x)$:

$$\frac{\#\{i \ : \ X_i = x\}}{n} \ \approx \ f(x)$$

- Example: Because CDF at x is $F(x) = \mathbb{P}(X \leq x)$:

$$\frac{\#\{i \ : \ X_i \leq x\}}{n} \ \approx \ F(x)$$

Frequency vs. probability: Proof

The relative frequency of B in the sequence can be written as

$$\frac{1}{n}\sum_{i=1}^{n} I_i, \quad \text{where} \quad I_i = \begin{cases} 1, & \text{if } X_i \in B, \\ 0, & \text{otherwise.} \end{cases}$$

I_i is the indicator variable for the event $\{X_i \in B\}$.

The random numbers $I_1, I_2, \dots$ are independent, and each has the same distribution as the first one I_1. (Why?)

By the law of large numbers, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} I_i \approx \mathbb{E}(I_1) = 0 \times \mathbb{P}(I_1 = 0) + 1 \times \mathbb{P}(I_1 = 1) = \mathbb{P}(X \in B).$$

Example: Empirical probabilities of dice

Trying to estimate $\mathbb{P}(X \leq 2)$, where X is a die result. This experiment is easy to repeat very many times, at least in simulation (random numbers in $\{1,2,3,4,5,6\}$ generated by computer).

n	est. probability	time
100	0.38000000	0.00 s
10000	0.33260000	0.00 s
1e+06	0.33351000	0.02 s
1e+08	0.33332494	1.55 s
1e+10	0.33333081	159.33 s

Here we know the true probability, so we see how the correct decimals increase (error decreases).

In reality we usually don't know the true probability, so we would like to estimate how big the error is. More about that later when we have more tools.

Empirical study of a probability

We can now empirically study the probability of an event, *if* we can repeat a similar experiment many times independently.

Question: Did we find the Holy Grail of probability calculus? We do not need cumbersome formulas, but for any event we just try many times and observe the relative frequency?

Partially true, but

- we need a method of performing the experiment many times (in reality or in a simulation)
- real-life repetitions could be difficult, expensive, dangerous
- simulation might (systematically) deviate from reality
- for large precision we need many repetitions: in fact, the error of our probability estimate is proportional to $1/\sqrt{n}$, so to get one more decimal place we need ... how many repetitions?

To add one more decimal place, we must cut the error to one tenth, requiring $100\times$ as many repetitions.

Using relative frequencies as empirical probabilities

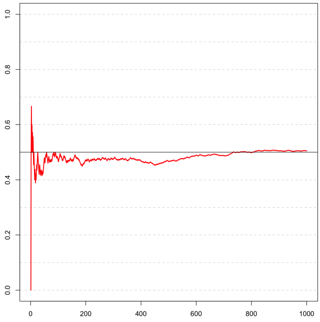
Still the fact, that relative frequencies in *long* sequences are fairly good estimates of probability, is the basis of much of modern statistics.

- sampling: we pick n persons from a population randomly; k of them have diabetes; guess that proportion k/n might be valid *in the population*
- clinical trial: we try a treatment n times, it works k times, we assume the same holds *in future treatments*
- an (empirical) histogram estimates a probability distribution
- Monte Carlo simulations in physics etc.: Simulate a process on computer millions of times and measure relative frequency. (Constructing the simulation might be the difficult part.)
- Monte Carlo integration: define a region in space, generate random points, see how often they land in the region $\rightarrow$ estimate the area of the region!

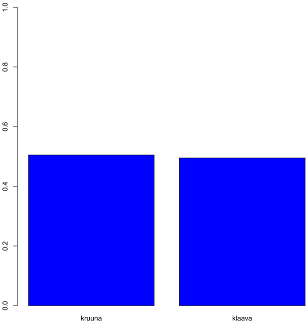
Example: 1000 coins

By LLN, relative frequency of *heads* in the random sequence $(X_1, \ldots, X_n)$ is

$$\frac{\#\{i \le n : X_i = \text{``heads''}\}}{n} \approx \frac{1}{2}$$



Relative frequency of heads as n grows

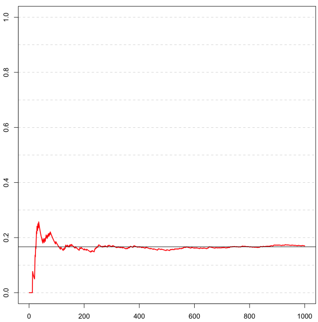


Relative frequencies of heads and tails in 1000 tosses

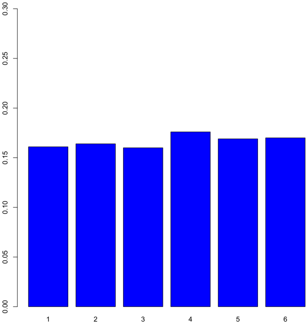
Example: 1000 dice

By LLN, relative frequency of *sixes* in random sequence $(X_1, \ldots, X_n)$ is

$$\frac{\#\{i \leq n : X_i = 6\}}{n} \approx \frac{1}{6}$$



Relative frequency of sixes as n grows



Relative frequencies of all six possible results in 1000 rolls

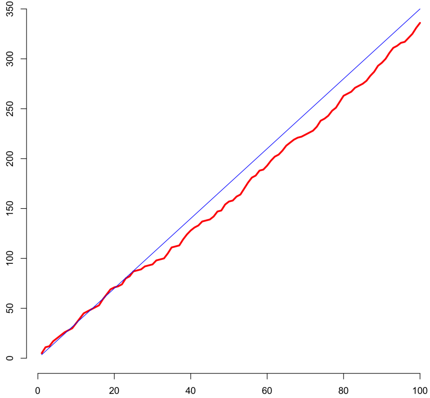
Example: Total payoff from dice

Suppose that on ith round, you get X_i euros if the result is X_i. Expected payoff from one round is $\mathbb{E}(X_i) = 3.5$ EUR.

By LLN, the *total* payoff from n rounds is *approximately*

$$\sum_{i=1}^{n} X_i = \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) n \approx 3.5n.$$

The red curve shows what actually happened (in one experiment).



Expected value vs. average: Summary

We have “average long-time” interpretations of both *expected value* and *probability*.

$$\mathbb{E}(X) \approx \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\mathbb{P}(X = x) \approx \frac{\#\{i \le n : X_i = x\}}{n},$$

where $X_1, X_2, \ldots$ are independent and identically distributed.

What if we *do not* have independent repetitions available?

- $X =$ next-year sales from a given startup company
- $X =$ next-year fire damages (if any) for a given house

Then $\mathbb{E}(X)$ still has some meaning, but “long-time average” might be difficult to realize.

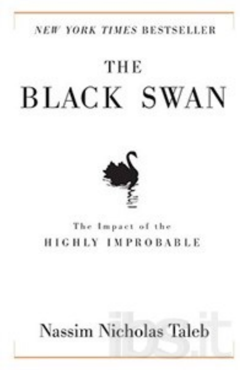
Example. "Black swan"

Consider the random variable distributed as

k	0	1000000
$\mathbb{P}(X = k)$	0.999999	0.000001

It has expected value

$$\mathbb{E}(X) = 0 \times 0.999999 + 1000000 \times 0.000001 = 1.$$



Now $\mathbb{E}(X) = 1$ tells something about the distribution, but not all.

If you generate independent random numbers from this distribution, the probability that the first 10 000 numbers *are all zeros*, is $0.999999^{10000} \approx 99\%$. After this observation, you might not expect anything else than zeros, but then...

http://www.fooledbyrandomness.com/

More about rolling rice

Zacariach Labby: Weldon's dice, automated
https://www.youtube.com/watch?v=95EErdouO2w
https:
//link.springer.com/article/10.1007/s00144-009-0036-8

Prof. Samuli Siltanen:
Samun tiedepläjäys: arpakuutio ja todennäköisyyden olemus
https://www.youtube.com/watch?v=rkJv4BveY4g
(in Finnish)

Tuomas Kukko & Risto Heikkinen:
Kimblen noppa ei ole täysin satunnainen
http://statistition.com/?p=440 (in Finnish)

Contents

Discretization of a continuous random variable

From a continuous variable X, we could make a new discrete variable $\lfloor X \rfloor_k = \frac{\lfloor 10^k X \rfloor}{10^k}$ by truncating to k decimals. For example $\lfloor 1.52793 \rfloor_3 = 1.527$.

$$\begin{aligned}
\mathbb{E}(\lfloor X \rfloor_k) &= \sum_{i=-\infty}^{\infty} \frac{i}{10^k} \mathbb{P}\left(\lfloor X \rfloor_k = \frac{i}{10^k} \right) \\
&= \sum_{i=-\infty}^{\infty} \frac{i}{10^k} \mathbb{P}\left(\frac{i}{10^k} \le X < \frac{i+1}{10^k} \right) \\
&= \sum_{i=-\infty}^{\infty} \frac{i}{10^k} \int_{\frac{i}{10^k}}^{\frac{i+1}{10^k}} f(x)dx \;=\; \int_{-\infty}^{\infty} \lfloor x \rfloor_k \, f(x)dx.
\end{aligned}$$

Because $\lfloor X \rfloor_k \to X$ as the precision $k \to \infty$, let use define

$$\mathbb{E}(X) = \lim_{k\to\infty} \mathbb{E}(\lfloor X \rfloor_k) \;=\; \lim_{k\to\infty} \int_{-\infty}^{\infty} \lfloor x \rfloor_k \, f(x)dx \;=\; \int_{-\infty}^{\infty} x \, f(x)dx.$$

Expected value of a continuous random variable

Expectation of a continuous X is defined as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\, f(x) dx.$$

In a continuous sense, it is the *density-weighted average* of the possible values.

Example (Metro waiting time)

If the waiting time X is uniformly distributed in $[0, 10]$, it has density

$$f(x) = \begin{cases} \frac{1}{10}, & x \in (0, 10), \\ 0, & \text{otherwise}, \end{cases}$$

and then the expectation is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\, f(x) dx = \int_{0}^{10} x\, \frac{1}{10} dx = 5.$$

Continuous expectation — Examples

Example (If density is a polynomial)

Suppose that the repairing time X of a printer, in hours, is a continuous r.v. with density $f(x) = 2x$, when $0 < x < 1$.

Thus X is always in the interval $[0, 1]$, but *more probably* at the higher end (where density is greater).

Calculate the expected value of X. (poll)

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\ f(x) dx = \int_0^1 x\ 2x\ dx = \int_0^1 2x^2 dx = 2/3.$$

Continous expectation — Examples

Example (If density is exponential)

Insects hit the windscreen randomly. The time between two hits is X, which has exponential distribution with rate parameter $\lambda = 1$ (insects per minute). The density is

$$f(x) = e^{-x}$$

for $x > 0$, and zero elsewhere.
Now the *expected* time between hits is (poll)

$$\mathbb{E}(X) \; = \; \int_{-\infty}^{\infty} x \, f(x) dx \; = \; \int_{0}^{\infty} x \; e^{-x} dx \; = \; 1.$$

(For calculating the integral, you need integration by parts.)

Long-run interpretation. How long does it take until your windscreen has collected 50 insects? By LLN, probably approximately 50 minutes (long-run average 1 insect per minute).

Expected value of random variable: Summary

Discrete

- Eg. uniform in $\{1, \ldots, 6\}$, binomial distribution, geometric distribution

$$\mathbb{P}(X \in A) = \sum_{i \in A} f(i)$$

$$\mathbb{E}(X) = \sum_{x} x\, f(x)$$

Continuous

- Eg. uniform in interval $[0, 10]$, normal distribution, exponential distribution

$$\mathbb{P}(X \in A) = \int_{A} f(x)\, dx$$

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\, f(x) dx$$

Contents

Example: Square of a discrete r.v. (directly)

Problem (Recall last lecture's square tile machine)

Calculate $\mathbb{E}(X^2)$, when X has distribution

k	0	1	2
$\mathbb{P}(X = k)$	0.2	0.5	0.3

Solution

$Y = X^2$ is discrete, with possible values $\{0, 1, 4\}$ and distribution

k	0	1	4
$\mathbb{P}(Y = k)$	0.2	0.5	0.3

Thus

$$\mathbb{E}(X^2) = \mathbb{E}(Y) = 0 \times 0.2 + 1 \times 0.5 + 4 \times 0.3 = 1.7.$$

Example: Cube of a continuous r.v. (directly)

Machine making cubes with side uniformly distributed in $[0, 10]$.

Problem

Calculate $\mathbb{E}(X^3)$, when X has uniform distribution in $[0, 10]$.

Solution

Define $Y = X^3$. It takes values $t \in [0, 1000]$. For those values,

$$F_Y(t) = \mathbb{P}(Y \le t) = \mathbb{P}(X^3 \le t) = \mathbb{P}(X \le t^{1/3}) = \frac{t^{1/3}}{10}.$$

and then we have density $f_Y(t) = \frac{t^{-2/3}}{30}$, thus

$$\mathbb{E}(X^3) = \mathbb{E}(Y) = \int_0^{1000} t \, \frac{t^{-2/3}}{30} dt = \frac{1}{30} \int_0^{1000} t^{1/3} dt$$

$$= \frac{1}{30} \times \left[\frac{3}{4} t^{4/3} \right]_0^{1000} = \frac{1000^{4/3}}{40} = 250.$$

Expectation of a transformed r.v. (Transformation formula)

If g is a function from the possible values of X into real numbers, then $g(X)$ is a random number; for each outcome s, this number becomes $g(X(s))$.

Fact

- *For a discrete random variable,*

$$\mathbb{E}(g(X)) = \sum_{x} g(x)\, f(x).$$

- *For a continuous random variable,*

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, f(x) dx.$$

Example: Square of a discrete r.v. (transformation formula)

Problem

Calculate $\mathbb{E}(X^2)$, when X has distribution

k	0	1	2
$\mathbb{P}(X = k)$	0.2	0.5	0.3

Solution

Apply the transformation formula with $g(k) = k^2$,

$$\mathbb{E}(X^2) \;=\; \sum_k k^2 f(k) \;=\; 0^2 \times 0.2 + 1^2 \times 0.5 + 2^2 \times 0.3 \;=\; 1.7.$$

Example: Cube of a continuous r.v. (transformation formula)

Problem

Calculate $\mathbb{E}(X^3)$, when X has uniform distribution in $[0, 10]$.

Solution

Apply the transformation formula with $g(t) = t^3$,

$$\mathbb{E}(X^3) = \int_{-\infty}^{\infty} t^3 f(t) dt = \int_0^{10} t^3 \frac{1}{10} dt = \frac{1}{10} \left[\frac{1}{4} t^4 \right]_0^{10} = 250.$$

This was much easier than with the direct method a few slides back.

Some easy transformations: Shifting and scaling

Lowercase letters are constants. Uppercase letters are random variables.

Fact

(i) $\mathbb{E}(a) = a$.

(ii) $\mathbb{E}(bX) = b\mathbb{E}(X)$.

(iii) $\mathbb{E}(a + bX) = a + b\mathbb{E}(X)$.

Proof.

(i) is obvious from definition of expectation.
(ii) If X is discrete, applying transformation $g(x) = bx$,

$$\mathbb{E}(bX) \;=\; \sum_x (bx) f(x) \;=\; b \sum_x x f(x) \;=\; b\mathbb{E}(X).$$

If X is continuous, similar proof (integrals instead of sums).
(iii) similarly by transformation $g(x) = x + a$ (whiteboard).

$\square$

Expectation from a multivariate function

Fact

- *For discrete random variables X and Y that have joint density $f(x, y)$,*

$$\mathbb{E}(g(X, Y)) \; = \; \sum_x \sum_y g(x, y) \, f(x, y).$$

- *For continuous random variables X and Y that have joint density $f(x, y)$,*

$$\mathbb{E}(g(X, Y)) \; = \; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \, f(x, y) dxdy.$$

Expectation from a multivariate function

Example (Box with two discrete dimensions)

A machine is making boxes whose bottom is a square with side X, and height is H. Thus the volume is $g(X, H) = X^2 H$.

The bottom side is 10 or 20, and height is 3 or 5, with joint density

	$H = 3$	$H = 5$
$X = 10$	0.4	0.3
$X = 20$	0.2	0.1

Expected value of volume is then

$$\begin{aligned}\mathbb{E}(g(X, H)) &= g(10, 3)f(10, 3) + g(10, 5)f(10, 5)\\ &\quad + g(20, 3)f(20, 3) + g(20, 5)f(20, 5)\\ &= (300 \times 0.4) + (500 \times 0.3) + (1200 \times 0.2) + (2000 \times 0.1)\\ &= 710.\end{aligned}$$

Contents

Sum of two random variables

Fact

$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Proof (discrete case).

Applying the multivariate transformation $g(x,y) = x + y$:

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_x \sum_y (x+y)\, f(x,y) \\
&= \sum_x \sum_y x\, f(x,y) + \sum_x \sum_y y\, f(x,y) \\
&= \sum_x x \left(\sum_y f(x,y) \right) + \sum_y y \left(\sum_x f(x,y) \right) \\
&= \sum_x x\, f_X(x) + \sum_y y\, f_Y(y) \\
&= \mathbb{E}(X) + \mathbb{E}(Y).
\end{aligned}$$

□

Sum of several random variables

For expectation of a longer sum, we can just apply $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ many times.

For a three-term sum $X + Y + Z$, observe that $X + Y$ is itself a random variable, we can call it U.

$$\begin{aligned}
\mathbb{E}(X+Y+Z) &= \mathbb{E}(U+Z) \\
&= \mathbb{E}(U) + \mathbb{E}(Z) \\
&= \mathbb{E}(X+Y) + \mathbb{E}(Z) \\
&= \mathbb{E}(X) + \mathbb{E}(Y) + \mathbb{E}(Z).
\end{aligned}$$

By the same method, we see that for any sum

$$\mathbb{E}(X_1 + \ldots + X_n) = \mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n).$$

So we can just take the expectations from each term separately. Together with scaling, this is known as linearity of expectation.

Example: Binomial distribution

Suppose we have n independent indicator variables $I_1, \ldots, I_n$, each indicating the success (1) or failure (0) of a random trial, with success probability $P(I_i = 1) = p$ and failure probability $q = 1 - p$. Then $X = \sum_{i=1}^{n} I_i$, the number of successes, has binomial distribution.
How to calculate $\mathbb{E}(X)$? You could try directly with $\sum_x xf(x)$, but it is difficult. Instead, take the expectation from each term separately.

$$
\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(I_1 + I_2 + \ldots + I_n) \\
&= \mathbb{E}(I_1) + \mathbb{E}(I_2) + \ldots + E(I_n) \\
&= p + p + \ldots p \\
&= np.
\end{aligned}
$$

E.g. $n = 100$ trials, $p = 0.20$ success probability $\implies$ expected value $np = 20$ successes.

You cannot move operations freely

We saw that *some* ("linear") operations can be "moved out" from inside the expectation, and vice versa:

- multiplication by a constant, $\mathbb{E}(bX) = b\,\mathbb{E}(X)$,
- addition of a constant, $\mathbb{E}(X + a) = \mathbb{E}(X) + a$,
- addition of two random variables, $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

This is *not generally true* for any operation you wish!

Example

The cube-making machine, with X uniform in $[0, 10]$. We calculated that $\mathbb{E}(X^3) = 250$.

However, $(\mathbb{E}(X))^3 = 5^3 = 125 \neq \mathbb{E}(X^3)$.

(Cube of expected value is not expected value of cube.)

Summary

The expected value $\mathbb{E}(X)$ is an *approximation* of the *average* of a large number of independent random numbers that are distributed the same as X.

Discrete

$$\mathbb{E}(X) = \sum_x x\, f(x)$$

$$\mathbb{E}(g(X)) = \sum_x g(x)\, f(x)$$

Continuous

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\, f(x) dx$$

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, f(x) dx$$

$$\mathbb{E}\left(a + \sum_{i=1}^{n} b_i X_i\right) = a + \sum_{i=1}^{n} b_i \mathbb{E}(X_i)$$

Contents

Further example. St. Petersburg paradox

A casino offers a gamble where you toss a coin repeatedly *until* heads. You gain

- 2 EUR, if heads occurs on 1st toss
- 4 EUR, if heads occurs on 2nd toss
- 8 EUR, if heads occurs on 3rd toss
- $\ldots 2^i$ EUR if heads occurs on ith toss ...

How much are you willing to pay, to play this game?

The payoff is a random number $g(T) = 2^T$, where game length T has discrete (geometric) distribution with density
$f_T(k) = (1/2)^k, k = 1, 2, 3, \ldots$
The *expected* payoff is

$$\mathbb{E}[g(T)] \;=\; 2^1(1/2)^1 + 2^2(1/2)^2 + 2^3(1/2)^3 + \cdots \;=\; \infty.$$

https://en.wikipedia.org/wiki/St._Petersburg_paradox

*Further exercise (outside required course)

Y = waiting time (minutes) if metros arrive at 10 min intervals, and stay 1 min.

This mixed distribution has (see previous lecture slides) CDF

$$F_Y(t) = \begin{cases} 0, & t < 0, \\ \frac{1}{10} + \frac{t}{10}, & 0 \leq t \leq 9, \\ 1, & t > 9. \end{cases}$$

Problem

Develop a meaningful definition for the expectation of a discrete-continuous mixed distribution, and calculate $\mathbb{E}(X)$.

Next lecture is about standard deviation and correlation. . .